A Magyar nemzeti szövegtár

Oravecz Csaba[1]

[1]Westpole Luxembourg
oravecz.csaba@gmail.com

1. A kezdetek[1]

A számítógépek megjelenésével szinte egy időben felmerült azok alkalmazása a nyelv elemzésére. Már a 1960-as években, de különösen a 80-as évektől a számítógépek tömeges elterjedésével új távlatok nyíltak a nyelv empirikus vizsgálata terén. Egy új ága született a nyelvészetnek, a korpusznyelvészet, amely a nyelvhasználat számítógépes modellezését tűzte ki célul. Ehhez nagy méretű szöveges adatbázisokat (korpuszokat) építettek, amelyek egy időben és térben meghatározott közösség nyelvhasználatának nyelvileg elemzett reprezentatív mintáját jelentik. Alapvetően a Nyelvtudományi Intézetben Váradi Tamás vezetésével 1997-ben létrejött, akkor még Korpusznyelvészetinek nevezett osztály is egy ilyen adatbázis, a Magyar nemzeti szövegtár (MNSz.) megépítését tűzte ki célul. Ez a kezdeményezés akkoriban nálunk mind módszertanában, mind volumenében újdonságnak számított (ez sokáig az osztályon dolgozó kevesek létszámában is megmutatkozott), de talán nem túlzás azt állítani, hogy a magyar nyelvre irányuló adatközpontú számítógépes nyelvészeti kutatásoknak és a nyelvfeldolgozó alkalmazások fejlesztésének egyik megkerülhetetlen kiindulópontjává vált. Ma már persze természetes a nagy mennyiségű nyelvi adat nélkülözhetetlensége, és a felhasználására vonatkozó egyre növekvő igény, de mintegy negyedszázaddal ezelőtt egy ilyen vállalkozás sok mindenben úttörőnek számított magyar nyelvterületen.

Az adatközpontú módszerek és alkalmazások fejlődése az utóbbi időben még inkább szembetűnő. A számítógépek teljesítményének növekedése, a szövegadatbázisokból, korpuszokból automatikusan tanulni képes, gépi tanuló eljárások kifejlesztése, a neurális hálózatok térhódítása

[1] Ez a dolgozat a Magyar Tudományban 2014-ben megjelent Váradi Tamás – Oravecz Csaba „A Magyar Nemzeti Szövegtár egymilliárd szavas új változata” című tanulmány alapján készült, abból kisebb változtatásokkal vagy anélkül hosszabb szövegrészeket is tartalmaz.

DOI: https://doi.org/10.18135/VT70.5

a számítógépes nyelvészet, a nyelvtechnológia intenzív fejlődéséhez vezetett, és tovább szélesíti, alakítja át a(z elméleti) nyelvészeti kutatások spektrumát és módszertanát is.

A számítógéppel végzett, illetve segített nyelvészeti kutatások, a nyelvtechnológiai alkalmazások alapvető feltétele a naprakész, a nyelvhasználatot reprezentatív módon tükröző, géppel olvasható és feldolgozható nyelvi adat, mely minden elméleti és alkalmazott kutatás kiindulópontja, a nyelvtechnológiai alkalmazások fejlesztésének nélkülözhetetlen nyersanyaga. Az ezeket az adatokat tartalmazó nyelvi adatbázisok pontos, számszerűsíthető képet adnak a nyelvhasználatról, egyben megkerülhetetlen forrásai és bemeneti adatai nyelvfeldolgozó algoritmusoknak, valamint értékes információt hordoznak az adott nyelvhez kötődő kultúra kutatóinak, társadalomtudósainak számára is.

A Szövegtár első változata 1998 és 2001 között készült, és a 90-es évek második felének nyelvhasználatából merített reprezentatív mintával a magyar nyelv első, az akkori gyakorlatban is jelentős méretűnek számító, nyelvileg elemzett korpusza volt, amely hálózati lekérdező felületen bárki számára szabadon hozzáférhető volt (Váradi, 2002). A munkálatok kezdetétől számított, lassan 15 év múltával vált nyilvánvalóvá, hogy általában a számítógépes korpuszokkal, így az MNSz.-szel szemben támasztott igények jelentős mértékben változtak és több szempontból megnövekedtek, különösen az alábbi 3 területen:

- **Minőség.** A számítógépes nyelvészeti technológia gyors fejlődése miatt az MNSz.-ben alkalmazott számítógépes nyelvi elemzés technológiája, pontossága és a nyelvi információ (re)prezentációjának módszere elmaradt a nemzetközi sztenderdnek tekinthető szinttől.
- **Terjedelem.** Az első változatban előirányzott 100 millió szavas terjedelem már nem volt jelentősnek tekinthető. Az adatközpontú módszerek/alkalmazások elterjedése és sikeressége a számítógépes nyelvfeldolgozás területén a nemzetközi gyakorlatban kívánatossá tették a milliárd szavas nagyságrendű korpuszok kifejlesztését (Parker et al., 2011), mivel az adatok ugrásszerű növekedése a rajtuk alapuló alkalmazások minőségének javulását vonja (vonta) maga után.
- **Reprezentativitás, lefedettség.** A nyelvhasználat pontos, akár a nyelvtörténeti kutatások igényeit is kielégítő dokumentálása egy-

részt újabb és újabb állapotfelvételt (adatgyűjtést) igényel, másrészt a nyelvi változatok széles skáláját kell, hogy képviselje. Ebből a szempontból például az MNSz. kritikus hiányossága volt a beszélt nyelvi adatok teljes hiánya.

2. Az MNSz. előzményei és első változata

Külföldi kutatások eredményeként már a 60-as évektől rendelkezésre állnak mai mértékkel természetesen kis méretűnek számító, de gondosan összeállított korpuszok (lásd Brown-korpusz [Kučera és Francis, 1967]). A 90-es évek jelentős produktuma az MNSz.-nek is néhány szempontból mintául szolgáló British National Corpus, és ettől az időszaktól folyamatosan készültek további nemzeti korpuszok. A nagy méretű korpuszokban reprezentált nyelvi információ gépi előállítására irányuló kutatás gyakorlatilag egy külön számítógépes nyelvészeti „iparág", a különféle annotáló és egyértelműsítő rendszerek fejlesztésének kialakulásához vezetett. Ebbe a sorba illeszkedett az MNSz. első változata is, ami az ilyen nagyságrendű korpuszokhoz hasonlóan automatikus morfoszintaktikai annotációt kapott a Nyelvtudományi Intézetben kifejlesztett nemzetközi szinten is élvonalbeli pontossággal működő eljárás segítségével (Oravecz és Dienes, 2002).

Az MNSz. első változatának elkészülte óta mind a korpuszok mérete, mind az alkalmazott gépi feldolgozás minősége és részletessége megváltozott. Ma már az elvárt kategóriát jelentik a több száz millió vagy több milliárd szavas adatbázisok, és ebben a nagyságrendben az MNsz.-szel párhuzamosan elkészült a magyar Webkorpusz is (Halácsy et al., 2003). A feldolgozás tekintetében hatékonyabb, pontosabb és részletesebb nyelvi elemzést adó eljárások, alkalmazások kifejlesztését célzó kutatások szintén a 2000-es évek elején-közepén kezdődtek meg magyar nyelvre is (Halácsy et al., 2006, 2007; Trón et al., 2005).

A jelentős méretű korpuszokban tárolt nyelvi elemzés részletessége automatikus annotáció esetén általában a morfológia szintjén maradt, szintaktikailag elemzett, magyar nyelven is létező adatbázisok az elfogadható elemzési pontosság érdekében (géppel segített) kézi annotációval készültek (Csendes et al., 2004).

Napjaink milliárd szavas szövegkorpuszai ún. opportunista összeállítással, általában a weben elérhető szövegek teljes letöltésével készülnek, azaz összetételükben kevésbé törekednek a nyelvhasználat különféle változatainak kiegyensúlyozott reprezentálására.

3. A továbblépés

Az MNSz. első változata igen sikeres nyelvi erőforrásnak volt tekinthető. A Kárpát-medencei Magyar Nyelvi Korpusz projekt keretében 2005 novemberére a határon túli nyelvváltozatokkal 187 millió szóra kibővült korpusznak több ezer regisztrált felhasználója volt, az MNSz.-ben található nyelvi adatok alapján több tucat tanulmány készült. Mindezek ellenére kétségtelen, hogy a mintegy 15 év elteltével az első változat elavulttá vált.

Az új változat (MNSz.^{2}) kifejlesztésének célja az 1. fejezetben említett hiányosságok kiküszöbölésével olyan magas minőségű, megnövelt és lefedettségét illetően kibővített komplex nyelvi adatbázis létrehozása volt, amely hatékonyan képes kiszolgálni a felhasználók és kutatók megnövekedett igényeit. Ennek érdekében a fenti felosztás szerint a új változattal kapcsolatos célkitűzések az alábbiakban foglalhatók össze:

- **Minőség.** A korpusz anyagának minden feldolgozási és elemzési lépésében új, korszerű számítógépes nyelvészeti technológia felhasználása a legújabb vonatkozó fejlesztéseinek figyelembevételével és a magyar nyelvre való alkalmazásukra irányuló célzott kutatással.
- **Terjedelem.** A korpusz anyagának bővítése minimum 1000 millió szóra.
- **Reprezentativitás, lefedettség.** Újabb mintavétel a mai magyar nyelvhasználatnak a Szövegtárban addig is szereplő, valamint további változataiból. Jelentős hozzáadott értékként jelent meg a beszélt nyelvi megnyilatkozások lejegyzett formátumát tartalmazó korpuszrész kialakítása, valamint mintavétel a közösségi média szövegeiből.

4. Az új változat fejlesztése

Az MNSz.^{2} esetében az MNSz. első változatában alkalmazott technológia minden részletében felülvizsgálatra, átdolgozásra, továbbfejlesztésre került a nemzetközi eredmények és a magyar nyelvre irányuló újabb kutatások alkalmazásával. Ez a munka a korpuszépítés minden fázisában jelentkezett.

4.1. Az anyaggyűjtés

Szöveges adatok összegyűjtésére ebben a nagyságrendben a kézenfekvő módszer vagy az internet bizonyos tartományainak végigpásztázása és

az ott talált anyagok valamilyen heurisztikus szűréssel segített, de alapjában véve válogatás nélküli letöltése, vagy nagy mennyiségű sajtóanyag beszerzése. Kizárólagos alkalmazás esetén mindkét módszernek vannak egyértelmű hiányosságai, ha a cél egy kiegyensúlyozott, elegendő metaadattal ellátott korpusz összeállítása. Előbbi módszer a szűrés ellenére is gyakran nagyon zajos adatot eredményez, melyhez jellemzően az az alapvető bibliográfiai információ is hiányzik, amely nélkül alapos nyelvészeti kutatások sokszor nemigen végezhetők. Az utóbbi módszerrel előálló korpusznak a reprezentativitás hiánya a szembetűnő hátránya.

Ezért jelentős munkát kellett fordítani a korpusz anyagának kontrollált és az adott forráshoz illeszkedő begyűjtésére: a közösségi médiából származó szövegek automatikus monitorozására, számítógéppel feldolgozható és metaadatolható eredményt adó letöltésére, a különböző forrásgazdákkal történő megegyezésre az általuk birtokolt anyagok archívumához való hozzáféréshez. Azok a források, melyek már alapesetben valamilyen (félig) strukturált, jól feldolgozható formátumban álltak rendelkezésre, előnyt élveztek a vegyes formátumú esetleges összeállítású archívumokkal szemben. A gyűjtés nagyságrendje természetesen eleve kizárta a kézi beavatkozást és a nagyon zajos kimenetet adó módszereket, mint a dokumentumok szkennelése, illetve optikai karakterfelismerést igénylő dokumentumok felhasználása. Az a manuális munkaerő, amely ezeket a módszereket alkalmazhatóvá tette volna, messze nem állt rendelkezésre.

Az anyaggyűjtés során elkerülhetetlenül szembesülni kellett az utóbbi időben egyre nagyobb hangsúlyt kapó szerzői jogokkal kapcsolatos kérdésekkel. Ekkora nagyságrendben lehetetlen vállalkozás volt minden adatgazdától (ha egyáltalán beazonosítható és megtalálható) a lehető legszabadabb felhasználói jogok megszerzése. Az MNSz.^{2} így alapesetben továbbra is egy felhasználói felületen férhető hozzá.

Az az előzetes várakozás, hogy a 15 évvel ezelőtti helyzethez képest a szöveges dokumentumok kezelése és tárolása a nemzetközi szabványokhoz közelítve sokat javult, és ez majd nagyban megkönnyíti a korpusz anyagának összegyűjtését, nem igazolódott be; sok probléma adódott a forrásszövegek hozzáférhetőségével és eredeti formátumával. Ehhez adódott még egy sajnálatos további hátráltató tényező: számos olyan adatforrás, amelyeknek a szövegei az MNSz. első változatának szerves részét alkotják, nem járult hozzá az azóta keletkezett szövegeik felvételéhez az MNSz.^{2}-be. Ennek valódi okait csak találgatni lehet, szomorú következménye viszont az, hogy a nyelvhasználat bizonyos jelentős szegmentumai az MNSz.^{2} mintavételéből teljesen kimaradtak.

A korpusz végül mintegy 1,5 milliárd szóra bővült. A sajtónyelvi anyag továbbra is domináns maradt, viszont minden nyelvváltozat anyaga minimum megduplázódott a korábbi változathoz képest, valamint megjelent egy új „műfaj", a(z átírt) beszélt nyelvi anyag is.

4.2. Előfeldolgozás és szövegnormalizálás

Az előfeldolgozás és normalizálás során a cél a forrásszövegek olyan szabványos elektronikus formátumba alakítása volt, mely hatékonyan feldolgozható bemenetként szolgálhat a nyelvi elemzőlánc számára. Ebben a lépésben történik a forrásformátumokból a hasznos szöveges tartalom kinyerése és az alapvető dokumentumstruktúra azonosítása, a karakterek normalizálása. A későbbi feldolgozás szempontjából fontos lépés a nyelvazonosítás, a nem magyar nyelvű szövegrészek kiszűrése, illetve megjelölése.

A gondos forrásválogatás ellenére a szövegek között mindig megjelennek (közel) duplikátumok. Ezek detektálása az MNSz.[2] esetében annál komplexebb kérdésnek bizonyult, hogy például egy, az internetről letöltött szövegeken alapuló korpuszokra kifejlesztett sztenderd megoldást közvetlenül alkalmazni lehessen (Pomikalek, 2011). A források változatossága (a közösségi média letöltött szövegeitől a hivatalos, jogi anyagokon keresztül a sajtószövegekig és a szépirodalomig) célzott módszer alkalmazását tette szükségessé, ami egy általános eszközkészleten alapult (Kupietz, 2005), de az egyes szövegtípusokra szabott automatikus detektálást manuális ellenőrzésnek is kellett követnie, hogy megállapíthassuk, vajon valódi duplikátumokról van-e szó, vagy olyan ismétlődő szövegegységekről, melyek szerves tulajdonsága az ismétlődés, így adattorzítást éppen az eltávolításuk okozott volna (lásd például az időjárásjelentések szövegei).

4.3. Elemzés és annotáció

Az MNSz.[2] fejlesztése a nyelvi feldolgozás minden szintjén jelentős minőségi javulást eredményező új, illetve továbbfejlesztett eszközöket használt fel, többek között új automatikus egyértelműsítő architektúrát, illetve a kapott morfoszintaktikai elemzést reprezentáló új annotációs formátumot. Elsősorban a morfo(fono)lógiai és szintaktikai kutatások későbbi igényeinek figyelembevételével megvalósult a legkisebb azonosított alkotóelemek, az egyes morfémák reprezentálása, a főnévi csoportok és névelemek azonosítása; ezek az információk az MNSz.-ben még

nem voltak jelen, és ma is ritkaságnak számít ilyen méretű korpuszban a nyelvi információ ezen részletessége.

A hasznos szöveganyag nyelvi elemzésének előkészítő lépéseit (mondatokra, illetve szó jellegű elemekre bontás – szegmentálás/tokenizálás) a Huntoken eszköz továbbfejlesztett, „háziasított" változata végezte (Miháczi et al., 2003). A morfológiai elemzést, mely gazdag morfológiával rendelkező nyelvekre kritikus fontosságú a további magasabb szintű elemzéshez, a jelentősen felújított Humor morfológiai elemző (Prószéky és Tihanyi, 1996) szolgáltatta, információt adva a szótővel, egyes morfémákkal, szóösszetételekkel kapcsolatban.

A belső annotációs formátum kiindulópontja a mondatra bontás és a tokenizálás kimenete. Minden szóelem *(token)* külön sorban szerepel, üres sorok jelölik a mondathatárokat. Minden további nyelvi annotáció típusonként egy-egy újabb oszlopban jelenik meg, egy rugalmas és könnyen feldogozható formátumot eredményezve. A több szóelemen átnyúló szerkezeteket az ún. IOB formátum szerinti kódolás[2] reprezentálja. Ez a belső reprezentáció egyszerűen átalakítható szabványos XML-formátumra, amennyiben szükséges.

5. Közzététel

Az adatbázis kialakításának utolsó lépéseként a megnövelt terjedelem igényelte az adatbázist építő rendszer továbbfejlesztését is. A megnövekedett felhasználói igények kiszolgálására az MNSz.[2] teljesen új hálózati felületet kapott, a lekérdezések beépített elemzését és több szempontú rendezését segítő korszerű webes technológiát kihasználó segédeszközökkel. A felület lehetőséget ad összetett menüvezérelt keresésre a kódolt információ minden részletében. A megjelenítési beállításokban a szövegkörnyezet, a metaadatok prezentációja állítható be, a kapott adatokon pedig további feldolgozási lépések végezhetők el, mint például megoszlásvizsgálatok, többszintű gyakorisági listák, többszavas kifejezések, kollokációk, igei argumentumok kinyerése.

6. Összegzés

Az MNSz. hivatkozási és látogatottsági adatai alapján egyértelmű, hogy az adatbázis a mai napig megkerülhetetlen forrása minden olyan kutatás-

[2] Inside, Outside, Beginning: szerkezet kezdő, szerkezeten belüli, szerkezeten kívüli elem.

nak és fejlesztésnek, amely magyar nyelvi adatot használ fel. A Szövegtár létrehozásával foglalkozó projekt hosszú időn keresztül a Korpusznyelvészetiből Nyelvtechnológiaivá vált osztály, de egyben a Nyelvtudományi Intézet zászlóshajója volt. Váradi Tamásnak az általa megalapított és irányított osztály központi tevékenységével kapcsolatos, a 90-es évek végén megfogalmazott jövőképe teljes mértékben beigazolódott.

Bibliográfia

Csendes, D., Csirik, J., Gyimóthy, T.: The Szeged Corpus: A POS Tagged and Syntactically Annotated Hungarian Natural Language Corpus. In: Sojka, P., Pala, K., Kopecek, I. (szerk.) Text, Speech and Dialogue: 7th International Conference, TSD. pp. 41–47. Springer (2004)

Halácsy, P., Kornai, A., Németh, L., Rung, A., Szakadát, I., Trón, V.: A Szószablya projekt. In: Alexin Z., Csendes D. (szerk.) Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem (2003)

Halácsy, P., Kornai, A., Oravecz, Cs.: HunPos – an open source trigram tagger. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague. (2007)

Halácsy, P., Kornai, A., Oravecz, Cs., Trón, V., Varga, D.: Using a morphological analyzer in high precision POS tagging of Hungarian. In: Proceedings of LREC 2006, pp. 2245–2248. (2006)

Kupietz, M.: Near-Duplicate Detection in the IDS Corpora of Written German. Technical Report IDS-KT-2006-01, Institut für Deutsche Sprache (2005)

Kučera, H., Francis, W. N.: Computational Analysis of Present-Day American English. Brown University Press, Providence, RI. (1967)

Miháczi, A., Németh, L., Rácz, M.: Magyar szövegek természetes nyelvi előfeldolgozása. In: Alexin Z., Csendes D. (szerk.) Magyar Számítógépes Nyelvészeti Konferencia. pp. 38–43. Szegedi Tudományegyetem (2003)

Oravecz, Cs., Dienes, P.: Efficient stochastic part of speech tagging for Hungarian. In: Rodríguez, M. G., Suarez Araujo, C. P. (eds.) Proceedings of the Third International Conference on Language Resources and Evaluation. pp. 710–717. ELRA, Las Palmas (2002)

Parker, R., Graff, D., Kong, J., Chen, K., Maeda, K.: English Gigaword Fifth Edition. Linguistic Data Consortium. (2011)

Pomikalek, J.: Removing Boilerplate and Duplicate Content from Web Corpora. Doktori disszertáció, Masaryk University, Faculty of Informatics, Brno. (2011)

Prószéky, G., Tihanyi, L.: Humor – A morphological system for corpus analysis. In: Rettig, H. (ed.) Proceedings of the first TELRI seminar in Tihany. pp. 49–158. Budapest (1996)

Trón, V., Gyepesi, Gy., Halácsy, P., Kornai, A., Németh, L., Varga, D.: Hunmorph: open source word analysis. In: Proceedings of the ACL 2005 Workshop on Software. pp. 77–85. The Association for Computational Linguistics (2005)

Váradi, T.: The Hungarian National Corpus. In: Rodríguez, M. G., Suarez Araujo, C. P. (eds.) Proceedings of the Third International Conference on Language Resources and Evaluation. pp. 385–389. ELRA, Las Palmas (2002)